**Predicting Hospital Length of Stay**

**Names**

**Abstract:**

Increasing healthcare costs in an unstable political and social landscape in the United States requires better understanding of how costs arise. Patient length of stay (LoS) in a hospital is one factor that directly contributes to increased healthcare costs. Predictive models can be created to predict patient length of stay to better understand the factors the influence LoS that can be addressed by changes in management, administration and policy. Using a dataset on 3612 patients collected by Good Health Corporation, we produced a multiple linear predictive model for LoS based on X predictors with an adjusted R2 value of , a CP value of X and an AIC of X. The model was validated using a bootstrap method. As expected, predictive elements were all determinants of the patient’s previous health. However, due to the low R2 value, it is suggested that a different modeling technique be used to optimize predictive capabilities to better alter management and administration to reduce costs.

**Introduction**

High healthcare costs are of primary concern in the United States (Ravi B. Parikh, 2017). Costs are consistently rising with questionable improvements in quality of care or life in relation to other nations (Joseph L. Dieleman, 2016). Due to rising costs and systemic concerns, there has been widespread debate on healthcare policy and management to optimize the current system (Barack H. Obama, 2017) (Clinton, 2016) (Henry J. Aaron, 2017). To optimize management and healthcare administration, we need a better understanding of the factors that influence overall costs. One such factor: patient length of stay (LoS) in a hospital, is directly associated with cost (Paul A Taheri, 2000) (Fine MJ, 2000). There have been various attempts at predicting patient length of stay using statistical modelling (A. Azari, 2013) (Gordon H. Robinson, 1966). Good Health Corporation (GHC) has collected data on 3612 patients admitted into the hospital including LoS and other possible predictors and has requested that a predictive model with Los as the outcome be created based on input predictors. To better understand the factors that affect LoS so that these can be addressed and optimized to reduce healthcare costs, we created a predictive model using multiple linear regression, based on the GHC data, that predicts patient length of stay based on relevant predictor variables.

**Methods**

**Data Source**

GHC collected a total of 3682 visit records on 3612 patients who were admitted to the hospital and over the age of 17. The visit had to have occurred within 24 hours of hospital admission. Data collected for each visit included: length of stay in the hospital in days, the modified early warning score (MEWS), the Charlson Comorbidity Index rank, if the patient had an ICU visit during hospitalization, the number of ER visits in the previous 6 months, the patient’s insurance type, patient demographics, and patient vital signs.

**Data Processing and Cleaning**

The GHC dataset was processed and cleaned to ensure data accuracy and therefore model validity. If a patient had more than one visit, only the first visit was included for model building. Due to skew and nonnormality of the LoS outcome variable, LoS was natural logarithm transformed. For vital sign predictor variables, outliers were identified using the standard z-score method, where values outside of the middle 99.9% of the distribution, or 3.291 standard deviations away from the mean, were replaced with the mean for the predictor. For O2 saturation all values over 100% were removed. This processing removed unrealistic values such as temperatures over 50degrees Celsius. For model building, we only used relevant predictors including: 30 day readmit rate, ER visits in past 6 months, Charson Index rank, MEWS, ICU visit during hospitalization, demographics (age, race, marital status), insurance type and vital signs (respiration rate, O2 saturation, BMI, Heartrate, Temperature, diastolic blood pressure and systolic blood pressure). We decided to omit data on patient religion due to lack of relevance.

**Model Selection**

Using our selected predictors, we utilized both stepwise regression and criterion based automatic procedures to select the best multiple linear regression model. The final model with the highest adjusted R2 value, lowest CP value and as few predictors as possible to ensure usability was selected.

**Model Diagnostics**

Once the optimal model was selected, model assumptions were checked. A residuals vs fitted value plot was created to detect for error heteroscedasticity. A quantile-quantile plot was created to detect normality of residuals. A scale-location plot was created to detect residual spread. A residuals versus leverage plot was created to help identify influential cases. Model outliers in the LoS were screened for using studentized residuals. Outliers were removed, and we remodeled without the outliers. As the adjusted R2 value increased we kept the new model. Leverage values in predictors were screened for as were influential predictors and none were significant enough to remove. Finally, multicollinearity was also screened for using VIF values and the findings were not significant.

**Model Validation**

We validated our final model using a Bootstrap method using 1000 repeats and calculated bias estimates for out model coefficients.

**Results:**

**Data Summary**

The mean length of stay was 5.461 days with a standard deviation of 5.92. Data cleaning and missing data resulted in some loss of data for certain predictors as seen in the reduces n in table 1. Summary statistics for all relevant continuous predictors used in model selection and for length of stay are included in table 1 and proportions for categorical variables are included in table 2.

Table 1.

---------------------------------------------------------------------

variable n mean sd minimum maximum median

----------------- ------ ------- ------- --------- --------- --------

losdays2 3612 5.461 5.92 0.04167 87.96 3.833

ageyear 3612 65.69 18.69 18 105 68

evisit 3612 1.754 1.577 0 4 1

bmi 2929 28.35 7.991 3.1 122.7 27.1

bpsystolic 3607 130.6 16.72 88.78 194 129.2

o2sat 3609 97.86 4.908 80 236.5 97.59

temperature 3610 36.73 0.899 11.85 52.27 36.73

heartrate 3607 80.07 13 37.58 242.6 79.2

respirationrate 3609 18.2 2.633 12 67.72 17.76

bpdiastolic 3611 72.52 9.798 29.56 154.4 71.85

---------------------------------------------------------------------

Table 2.

|  |  |
| --- | --- |
| Variable | N (%) |
| **Gender** | |
| Male | 1660 |
| Female | 1952 |
| **Race** | |
| White | 2057 |
| Black | 772 |
| Asian | 249 |
| Native American (Alaskan) | 22 |
| Native American (Hawaiian/Pacific Islander) | 4 |
| Other | 508 |
| **Insurance** | |
| Medicare | 1425 |
| Medicaid | 166 |
| Private | 1987 |
| Missing Data | 34 |
| **Marital Status** | |
| Single | 951 |
| Married | 1607 |
| Widowed | 690 |
| Divorced | 235 |
| Separated | 51 |
| Civil Union | 1 |
| Missing Data | 77 |
| 30 Day Readmit Rate | 517 |

**Final Model**

We selected a final model given by criterion based automatic model selection that included: 30 day readmit rate, ER visits, C Index, Age, Respiration Rate, Heart Rate, temperature and Systolic Blood pressure as predictors. Model coefficients are included in table 3. The adjusted R2 for the model is X the CP score is X and the AIC values was X. All model diagnostic graphs are included in figure 1. As seen, all model assumptions were met. Bootstap validation produced bias values for all model coefficients which are included in table 3. Bias values were low and non-significant.

Table 3

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Coefficient Value | P-Value | Bootstrap Bias |
| (Intercept) | -6.6121133 | 3.98e-07 |  |
| is30dayreadmit | 0.1807683 | 2.80e-05 |  |
| evisit | 0.0669272 | 4.55e-12 |  |
| cindex | 0.0421502 | 5.93e-07 |  |
| ageyear | 0.0109857 | < 2e-16 |  |
| respirationrate | 0.0660580 | 8.66e-13 |  |
| heartrate | 0.0076618 | 1.66e-09 |  |
| temperature | 0.1616098 | 4.61e-06 |  |
| bpsystolic | -0.0058352 | 1.49e-10 |  |

Figure 1.

**Discussion**

We were able to produce a predictive model for hospital length of stay using 8 predictors. Our model used a log transformation, meaning that each coefficient in table 3 means X controlled for all other variables. Most predictors were related to vital signs or current health status. This suggests that as expected the length of stay in a hospital is associated with how healthy the patient is or if they have recently been to the hospital (ER visits and admission to a hospital within the past 30 days). Our model suggests that factors such as insurance type, race or marital status have negligible impact on hospital stay as would be expected. It is worth noting that are multiple linear model does not fit the data well looking at the low adjusted R2 value. Though bootstrap validation suggested that our model is valid based on the data we have, it may not be the best model possible. Utilizing other models such as different fits (quadratic, exponential, etc.) or possibly more advanced machine learning techniques (artificial neural nets, deep learning, etc.) on larger data sets may produce a better predictive model. Add mean inflating R2

# References

A. Azari, V. P. (2013). Predicting Hospital Length of Stay (PHLOS): A Multi-tiered Data Mining Approach. *IEEE Explore*.

Barack H. Obama, J. (2017). Repealing the ACA without a Replacement — The Risks to American Health Care. *The New England Journal of Medicine*, 376:297-299.

Clinton, H. (2016). My Vision for Universal, Quality, Affordable Health Care. *The New England Journal of Medicine*, 375:e36.

Fine MJ, P. H. (2000). Relation between length of hospital stay and costs of care for patients with community-acquired pneumonia. *The American Journal of Medicine*, 378-85.

Gordon H. Robinson, L. E. (1966). Prediction of Hospital Length of Stay. *Health Services Research*, 287–300.

Henry J. Aaron, P. (2017). Which Road to Universal Coverage? *The New England Journal of Medicine*, 377:2207-2209.

Joseph L. Dieleman, R. B. (2016). US Spending on Personal Health Care and Public Health, 1996-2013. *JAMA*, 316(24):2627-2646.

Paul A Taheri, D. A. (2000). Length of stay has minimal impact on the cost of hospital admission. *Journal of the American College of Surgeons*, 123-130.

Ravi B. Parikh, M. M. (2017). Getting Real about Health Care Costs — A Broader Approach to Cost Stewardship in Medical Education. *The New England Journal of Medicine*, 376:913-915.